Hayal Dargin

Final Report:

Self-Injury and Usage of Social Media Analysis

Problem Statement

According to the researchers there has been a gigantic increase in depression and anxiety for American teenagers which began right around between 2011 and 2013. The number of teenage girls out of 100,000 in this country who were admitted to hospital every year because they cut themselves or otherwise harm themselves, that number was pretty stable until around 2010, 2011, and then it began going way up. According to the recent researchers, it is up 62 percent for the older teen girls. It is up 189 percent for the preteen girls which is nearly triple. We also see the same pattern with the suicide. The older teen girls are 15 to 19 years old; they are up 70 percent compared to the first decade of this century. The preteen girls, who have very low rates to begin with they are up to 151 percent. And that pattern points to social media available on mobile devices around 2009.

Gen Z, the kids born after 1996 or so, those kids are the first generation in history that got on social media in middle school. The whole generation is more anxious, more fragile, more depressed. This research will be looking into the relationship between the growth of technology (social media), usage of social media and self-injury (who wants to kill themselves or harm themselves) numbers of adolescents and young adults aged 10-14 and 15-19.

By using CDC (Centers for Disease Control and Prevention) Self Harm and Statista Monthly Active Users by My Space data, I created a tool to help some related departments such as Department of Education, organizations such as WHO, and families in realization of impact of use of social media.

Main purpose of this project is going to be suggesting that use of social media has a high positive correlation with the adolescents and young adults injuring themselves and attempting a suicide. I propose that in order to reduce injury rates and suicide attempts there should be strict restrictions on the number of social media apps that teenagers aged between 10 to 19 have access to on their phones and on any other devices. Since social media is supporting the trend advertisements, young adults should be prevented from using social media under some restrictions.

Data Wrangling

The row data set from CDC (Centers for Disease Control and Prevention) Self Harm contained 5889 rows with 21 columns, and the row data set from Statista Monthly Active Users by My Space contained 142 rows with 3 columns.

While the main data set of Self Harm was usable, it needed serious cleaning and wrangling, the second data set of Social Media was pretty clean and shaped.
I started by looking at the most important features of my interest for this project. Since subsetting the data frame is providing better look and understanding of features, I decided to select the columns that I will be working on from the Self harm data set. Following the subsetting the data set a massive cleaning has been done by checking the data types, missing values, definition of variable values, renaming some columns.

The final shape of my data set of Self Harm was 2631 rows with 8 columns, and Social Media was 142 rows with 3 columns.

Exploratory Data Analysis

In exploratory analysis, I wanted to do a general to specific study. First, I wanted to see the difference of self-harm between male and female. Then I continued on gender and try to answer the question: What is the highest self-harm/injury rate of male and female depending on the distribution of this difference according to different race, age group and years. I looked at correlations and differences between every column and my target parameter of injuries. I found strong outcomes and visualized them by bar plots.

Figure 1 plots the number of injuries among female and male. My first finding as we can see from the figure below, females are more likely to harm themselves than males.

Figure1

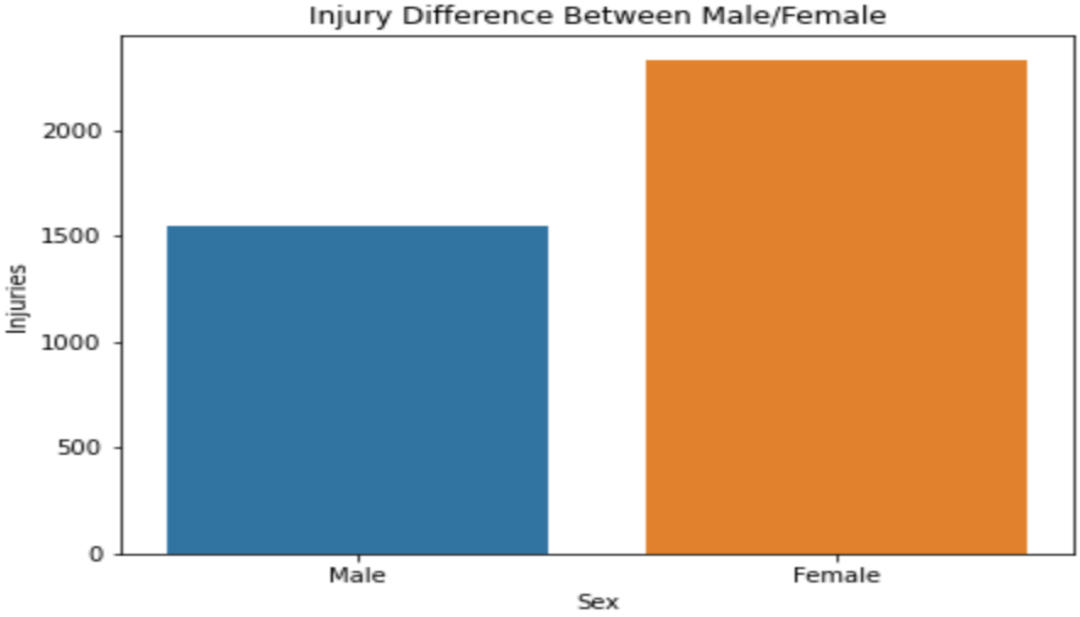


Figure 2 plots the number of injuries among different races and sex. Among all the races females are injuring themselves the most. Looking at the difference between races, female and male, white non-Hispanic females tend to injure themselves the most followed by white non-Hispanic males.

Figure2

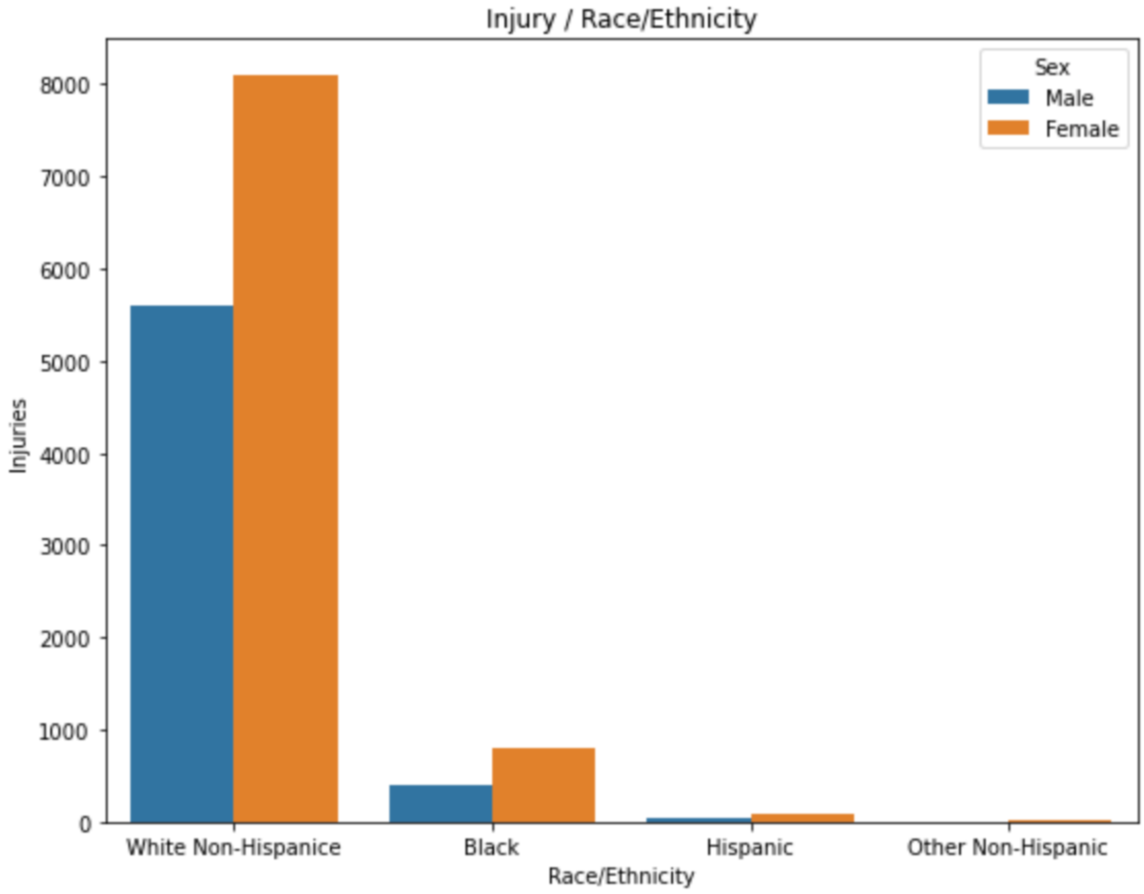


Figure 3 plots the number of injuries among different age groups and sex. In general, we see very clearly that the self-injury rate of women is always higher than men in all age groups. While there are no injuries between 0 and 9 years old, we see the highest number of injuries between 10-14 years and between 15-19 years old. We also see that at the age of 20, the injury numbers are arguably decreasing and continues in the same way.

The point I was trying to find was, first of all, which sex was most prone to self-injury. Then it was to find out in what age ranges there was the greatest difference. Up to this point, we see that White Non-Hispanic women and men between the ages of 10-14 and 15-19 have a very high rate of self-injury.

Figure3

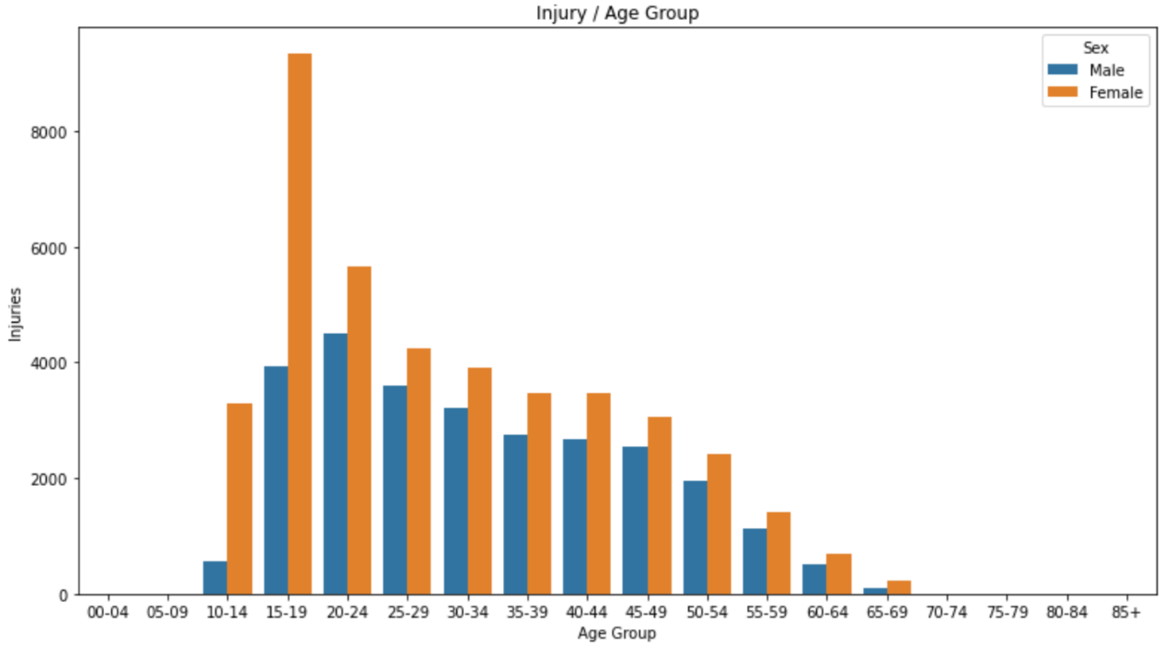


Figure 4 plots the number of injuries over the years with sex, and figure 5 plots the number of injuries over the years with Race and Ethnicity. In my data from 2001 to 2018, females always have a higher number of self-harm than male. While the number of self-harm is less in both female and male in the first 9 years, it has increased since 2010. Although there was a small decrease after 2013, the uniqueness started to increase as of 2017. As we can see in the figure 5, the increase in self-harm has not changed in White Non-Hispanic women and men followed by Black female and male. I wanted to recuperate, the number of self-harm among White Non-Hispanic female between the ages of 10 and

19 is higher than that of White Non-Hispanic male, and this rate of self-harm shows an increase in both males and females after 2010 in figure 6 below.

Figure4

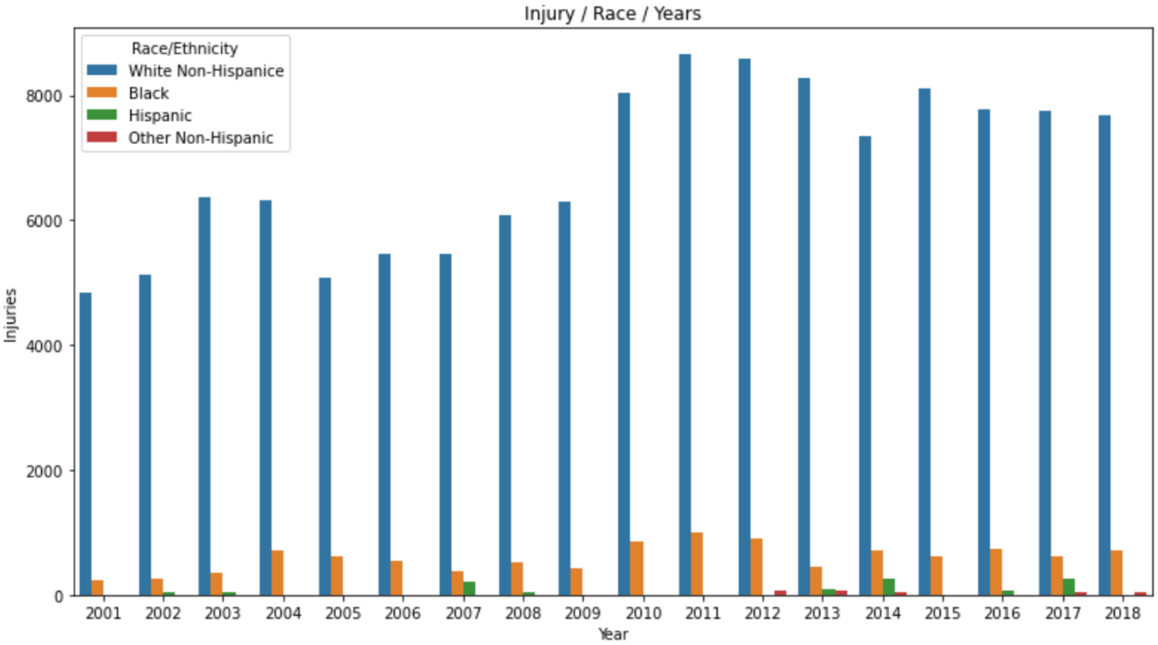


Figure5

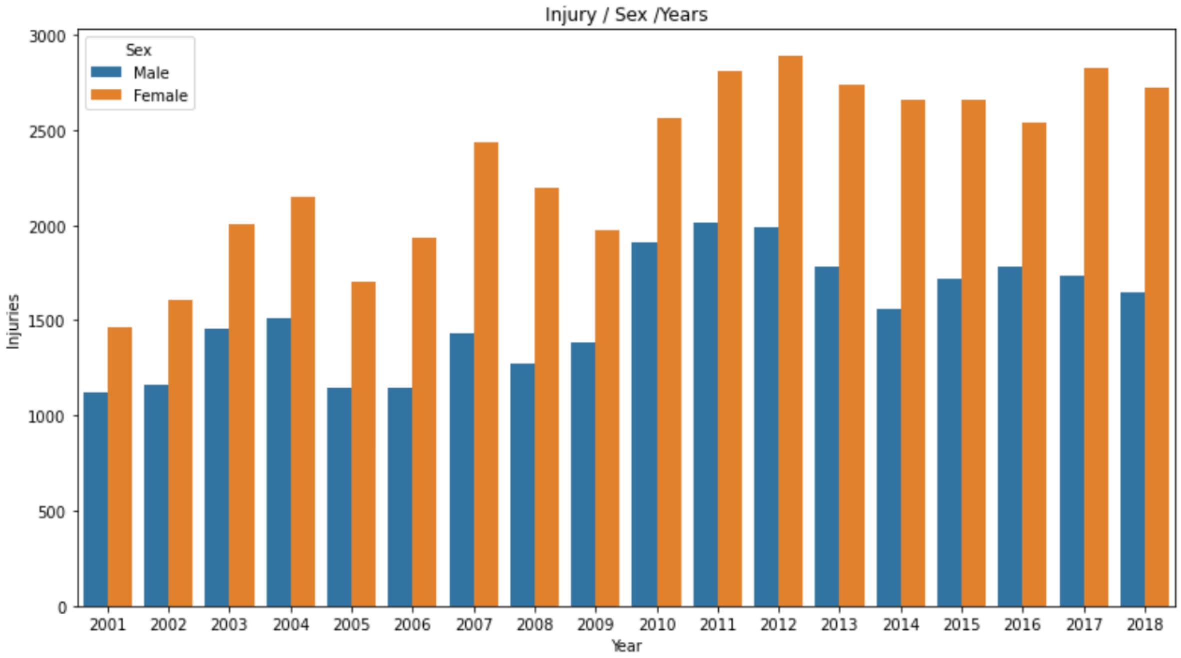


Figure6

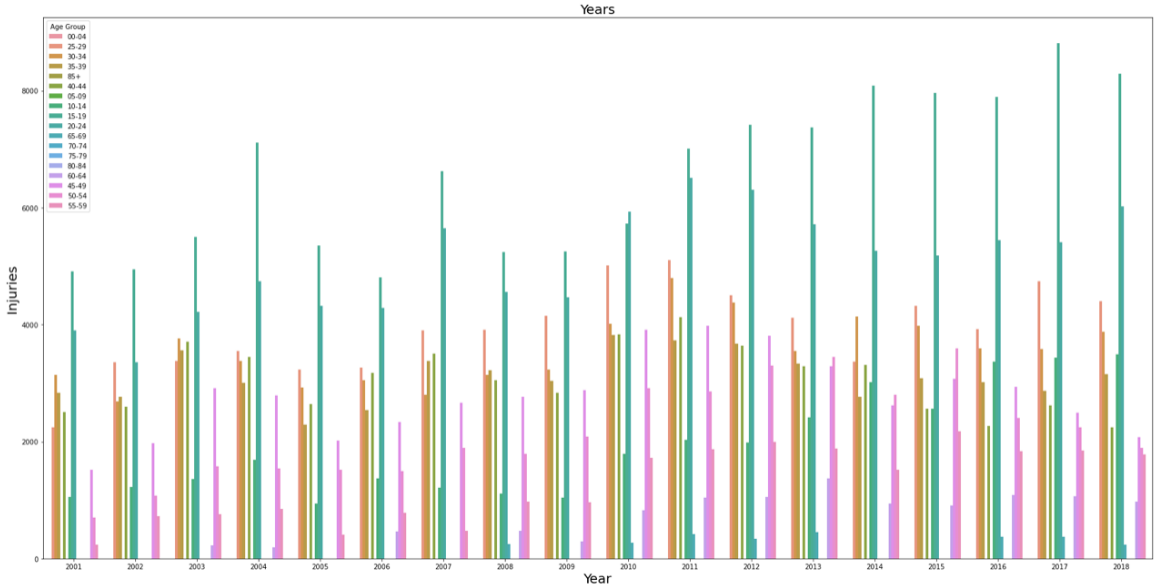
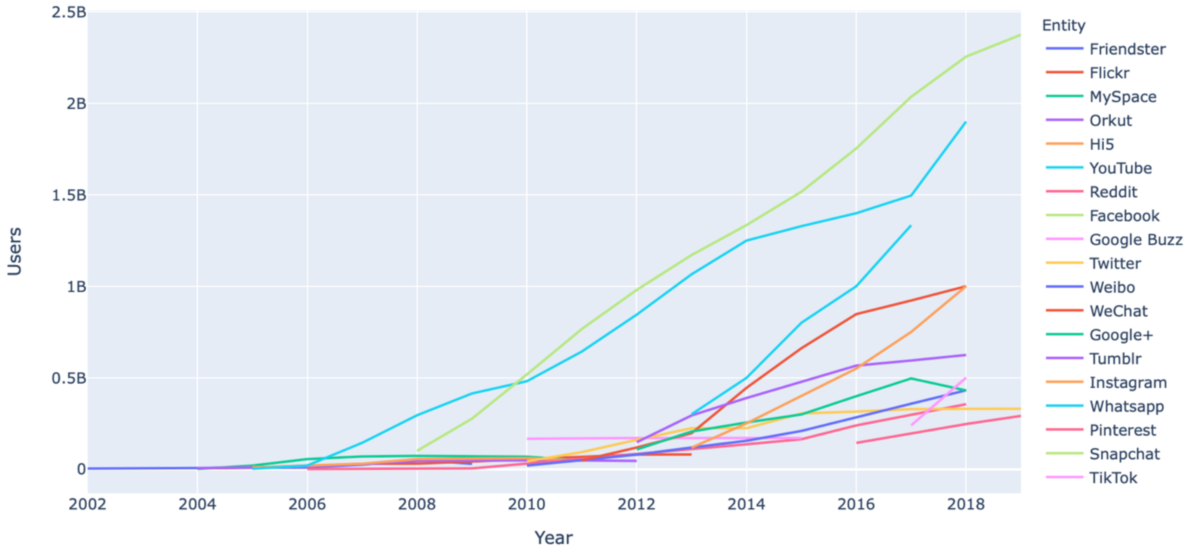


Figure 7 plots the number of social media users over the years with all the social media platforms. I ploted monthly active users, by platform from 2002 until 2018. We see that the first social media site to reach a million monthly active users was Friendster around 2002 followed by MySpace which achieved this milestone around 2004. Our plot shows that there are some large social media sites that have been around for ten or more years, such as Facebook, YouTube and Reddit; but other large sites are much newer such as TikTok, and Pinterest.

Figure7



The most controversial in social media data is the way the number of social media usage has taken over years. We see very clearly that while the number of social media platforms was less until 2010, the number of social media entities and the usage patterns increased rapidly after 2010.

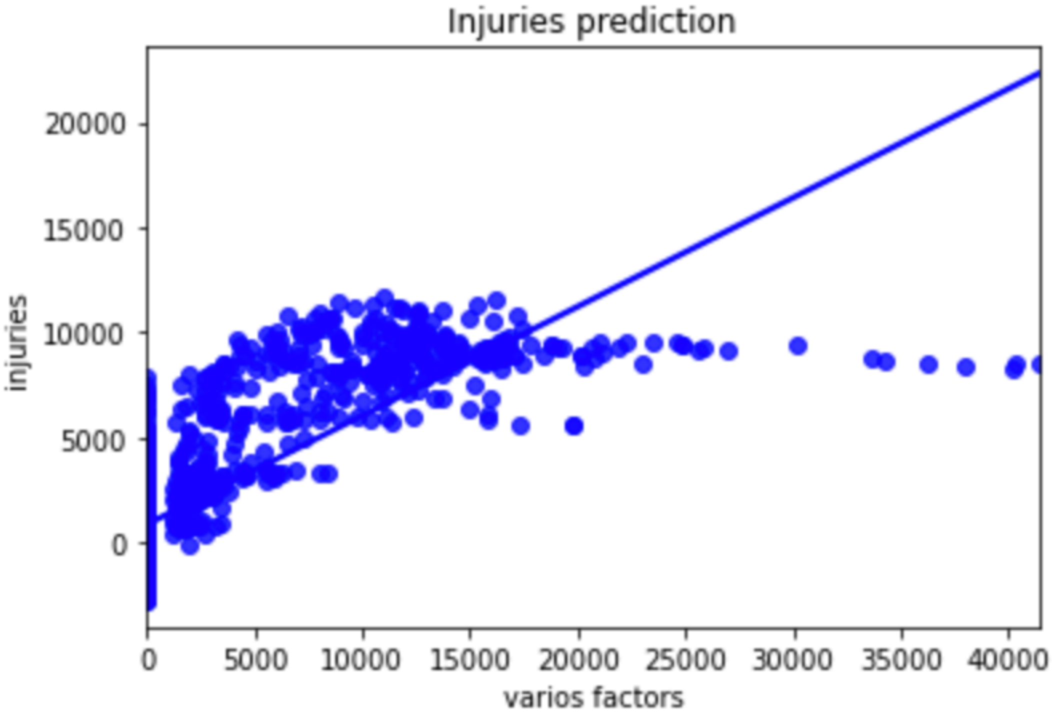
Model Selection

For modeling, I tested 2 different machine leaning regression models: Random Forest Regressor and Linear Regression. The metric I focused on when building my models were RMSE (root mean squared error), R Squared Score, MAE (mean absolute error). I wanted my models to predict features with high correlation and importance.

Before building my models, however, I needed to do some feature selection. To do this I dropped some unwanted columns. Then first I called Random Forest Regressor and passed in my training data to train the random forest regressor model. In order to select the best model, I evaluated my random forest model by rerunning the model with different numbers of trees and selected the best model with 30 n_estimatorts.

Next I build linear regression model and passed in my training data to train the linear regression model. Evaluated my models performance by using the RMSE(root mean squared error) and got higher RMSE on linear regression model as we can see in figure 8 below. This is because many values are zero in my data and there is very less linear correlation. In the future work we might want to eliminate these zero values to see better linear correlation. Thus, I used random forest regressor as my final model with lower RMSE.

Figure8



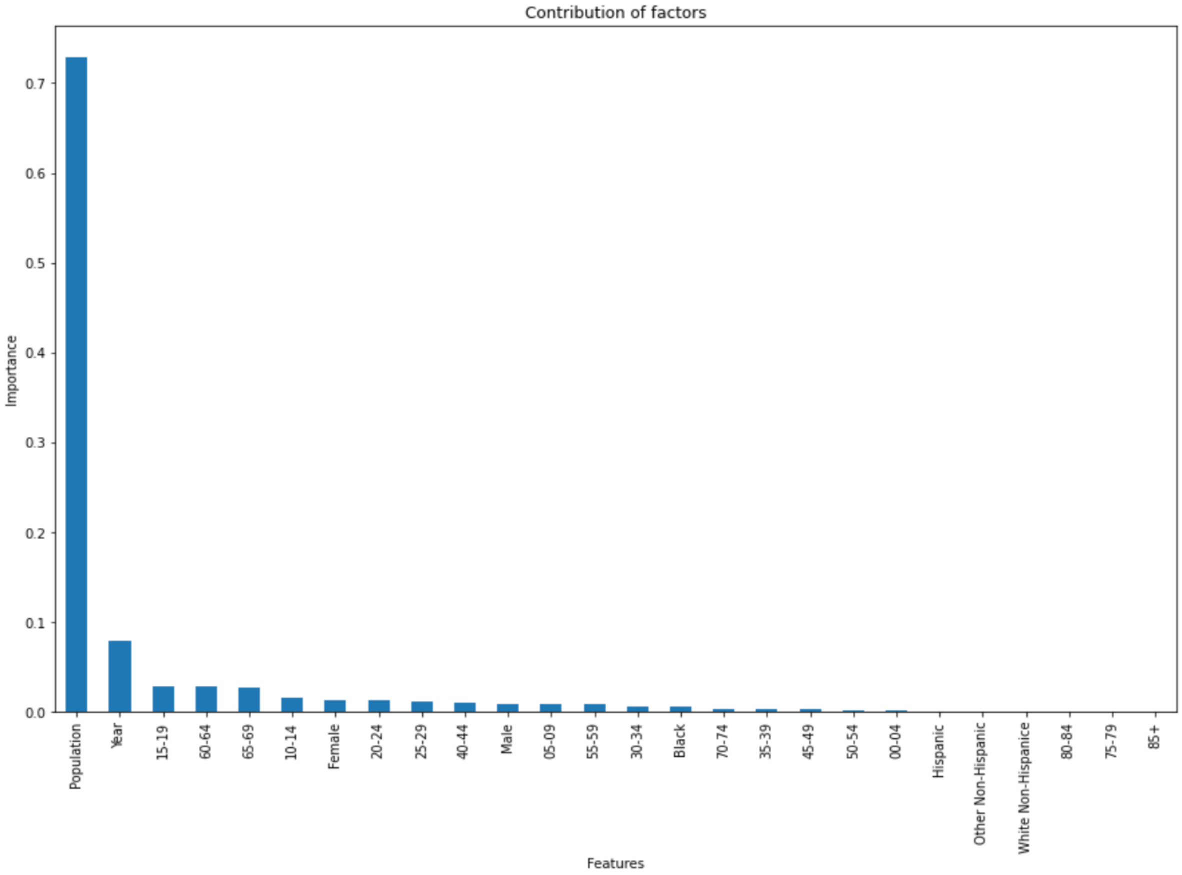
Conclusion

The data I am predicting in my model is injury numbers. In my model, I looked at the relationship of Injury variable with other variables such as age group, sex. The main hypothesis of my project is whether social media use is correlated with self-injury or not, if there is a correlation, which age groups, gender and ethnic identity.

From the figure 9 I can see that population and year have the highest degree of importance on injuries. Following, I see that the White Non-Hispanic people are more likely to commit self-harm followed by people who ages between 15-19 and 20-24.

The results I get from my analysis also support my model and analysis results. In my analysis, I found that self-injury is much higher in the 15-19 age group. When I look at the difference of self-injury numbers between men and women, I see that in all age groups the number of women is always higher than men.

Figure9



At the same time, when I look at the distribution of self-injury over the years, I found that this number starts to increase around the years 2009-2010. Likewise, the number of social media usage has increased around these years.

As a result, I predicted that the number of self-injuries of White Non-Hispanic people between the age group 15-19 and the age group 20-24 is higher than the other groups. Even more interesting is that this number is much higher in women.

Future Research

The biggest deficiency or difficulty I encountered in this project was the data I had. The fact that if I also had the age group, gender and ethnicity variables in my social media data set, which includes the numbers of social media usage, could give me much more meaningful results. Awareness of this can lead me to further steps in this project. With better data, one can look at the correlation of self-harm and social media use, and I believe the results will be much more meaningful.